Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region

Journal of General Virology (1993), **74**, 2391–2399. *Printed in Great Britain*

Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region

P. Simmonds,[1]* E. C. Holmes,[2]† T.-A. Cha,[3]‡ S.-W. Chan,[1] F. McOmish,[4] B. Irvine,[3] E. Beall,[3]§ P. L. Yap,[4] J. Kolberg[3] and M. S. Urdea[3]

[1] *Department of Medical Microbiology, Medical School, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG,* [2] *Division of Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JN, U.K.,* [3] *Chiron Corporation, 4560 Horton Street, Emeryville, California 94608, U.S.A. and* [4] *Edinburgh and South East Scotland Blood Transfusion Service, Royal Infirmary of Edinburgh, Lauriston Place, Edinburgh EH3 9HB, U.K.*

Hepatitis C virus (HCV) shows substantial nucleotide sequence diversity distributed throughout the viral genome, with many variants showing only 68 to 79% overall sequence similarity to one another. Phylogenetic analysis of nucleotide sequences derived from part of the gene encoding a non-structural protein (NS-5) has provided evidence for six major genotypes of HCV amongst a worldwide collection of 76 samples from HCV-infected blood donors and patients with chronic hepatitis. Many of these HCV types comprised a number of more closely related subtypes, leading to a current total of 11 genetically distinct viral populations. Phylogenetic analysis of other regions of the viral genome produced relationships between published sequences equivalent to those found in NS-5, apart from the more highly conserved 5′ non-coding region in which only the six major HCV types, but not subtypes, could be differentiated. A new nomenclature for HCV variants is proposed in this communication that reflects the two-tiered nature of sequence differences between different viral isolates. The scheme classifies all known HCV variants to date, and describes criteria that would enable new variants to be assigned within the classification as they are discovered.

Introduction

Infection with hepatitis C virus (HCV) has been identified as the major cause of post-tranfusion non-A, non-B hepatitis (Choo *et al.*, 1989; Kuo *et al.*, 1989). The virus has a positive-sense, ssRNA genome approximately 10 kb in length, with similarities in genome organization and some sequence homology with pestiviruses and flaviviruses (Choo *et al.*, 1991; Han *et al.*, 1991; Brown *et al.*, 1992). Different isolates of HCV show substantial nucleotide sequence variability distributed throughout the viral genome (Okamoto *et al.*, 1991, 1992*b*). Regions encoding the putative envelope protiens [E1, E2/non-structural protein 1 (NS-1)] are the most variable (Weiner *et al.*, 1991; Hijikata *et al.*, 1991), whereas the 5′ non-coding region (5′ NCR) is the most conserved (Han *et al.*, 1991; Cha *et al.*, 1991; Okamoto *et al.*, 1990; Bukh *et al.*, 1992*a*). Comparison of published sequences of HCV has led to the identification of a number of distinct virus 'types', that may differ from each other by as much as 33% over the whole viral genome (Choo *et al.*, 1991; Okamoto *et al.*, 1991; Chan *et al.*, 1992; Mori *et al.*, 1992; Okamoto *et al.*, 1992*b*).

This degree of sequence variability is sufficient to alter the antigenic and biological properties of members of this virus group significantly. The immunoreactive region of the NS-4 protein is highly variable so that most epitopes are type-specific (Simmonds *et al.*, 1993*b*). This leads to a substantial reduction in the effectiveness of antibody assays based on this protein for serological diagnosis of infection with divergent HCV types (Chan *et al.*, 1991; McOmish *et al.*, 1993; Kato *et al.*, 1991; Cha *et al.*, 1992); however, second generation assays containing multiple antigens are more broadly reactive. Samples found positive by second generation assays, but negative with first generation assays, have proved to be new viral types (Cha *et al.*, 1992). Whether or not other

† Present address: Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

‡ Present address: Aviron, Belmont, California, U.S.A.

§ Present address: Molecular Immunopathology Unit, Blood Transfusion Centre, Longlea Road, Cambridge CB2 2PT, U.K.

‖ Present address: Department of Biochemistry, University of California, Berkeley, California 94720, U.S.A.

The nucleotide sequences obtained in this investigation have been submitted to GenBank and assigned the accession numbers L23435 to L23475.

HCV variants remain undetected using the extensive battery of antigens currently employed remains an area of intensive investigation.

Variability in the envelope region is even greater so that neutralizing antibodies might be type-specific and allow multiple infection with different HCV variants in re-exposed individuals. Finally, there is some evidence for variation in the course of infection associated with different HCV variants and in response to treatment with interferon (Kanai *et al.*, 1992; Yoshioka *et al.*, 1992; Takada *et al.*, 1992*a*; Pozzato *et al.*, 1991). If these results are confirmed in larger studies this would indicate an important role for identification of genotype in the pretreatment assessment of patients with chronic hepatitis.

We have previously carried out phylogenetic analyses of nucleotide sequences amplified in the region of the genome encoding the core protein, and parts of the NS-3 and NS-5 proteins (Chan *et al.*, 1992). Although different degrees of variability were found in various parts of the genome, analysis of each produced trees topologically identical to those obtained upon analysis of complete genome sequences. Based on these initial results, we proposed that HCV might usefully be classified into three major HCV types, with variants designated type 1 and type 2 each comprising two more closely related subtypes (Chan *et al.*, 1992).

In this communication we have carried out phylogenetic analysis of a 222 bp fragment of NS-5 between nucleotide positions 7975 and 8196 (numbered as in Choo *et al.*, 1991). This region was amplified from plasma or serum of 41 infected individuals using relatively well conserved primers (Enomoto *et al.*, 1990), and compared with 35 published NS-5 sequences from other laboratories.

Methods

Samples. Plasma or serum samples were obtained from 41 HCV-infected blood donors or NANBH patients from a variety of geographical regions (Table 1). Sequences obtained in this study were compared with 35 previously published sequences listed in Table 1.

Nucleotide sequence analysis. To obtain sequences in the NS-5 region, viral RNA was reverse transcribed and amplified in a single reaction using primers thought to be highly conserved amongst different variants of HCV (Enomoto *et al.*, 1990). For some sequences, a second PCR was carried out with primers 554 and 555 (Chan *et al.*, 1992) in combination with two new primers, 122 (sense orientation; 5′ CTC AAC CGT CAC TGA GAG AGA CAT 3′) and 123 (antisense; 5′ GCT CTC AGG TTC CGC TCG TCC TCC 3′). Product DNA was phosphorylated, purified and cloned into the *Sma*I site of pUC19 (Yanisch-Perron *et al.*, 1985) following the procedures described elsewhere (Simmonds & Chan, 1993). Alternatively, amplified DNA was purified and directly sequenced as previously described (Simmonds *et al.*, 1990; Cha *et al.*, 1992). These methods allowed comparison of a 222 bp fragment of DNA homologous to positions 7975 to 8196 in the prototype virus (numbered as in Choo *et al.*, 1991).

Table 1. *Origin of HCV variants analysed in this study*

No.	Name	Origin	Reference
1	HCV-1	U.S.A.	(Choo *et al.*, 1991)
2	GM2	Germany	This paper
3	I21	Italy	This paper
4	SP2	Spain	This paper
5	US17	U.S.A.	This paper
6	HCV-H	U.S.A.	(Inchauspe *et al.*, 1991)
7	PT-1	Japan	(Enomoto *et al.*, 1990)
8	H77	U.S.A.	(Ogata *et al.*, 1991)
9	H90	U.S.A.	(Ogata *et al.*, 1991)
10	T1801	Scotland	This paper
11	T1825	Scotland	This paper
12	T2138	Scotland	This paper
13	GH6	Germany	This paper
14	JH	Japan	(Kubo *et al.*, 1989)
15	SP1	Spain	This paper
16	SP3	Spain	This paper
17	HCV-J	Japan	(Kato *et al.*, 1990)
18	HCV-BK	Japan	(Takamizawa *et al.*, 1991)
19	T	Taiwan	(Chen *et al.*, 1992)
20	K1	Japan	(Enomoto *et al.*, 1990)
21	K1-1	Japan	(Enomoto *et al.*, 1990)
22	K1-2	Japan	(Enomoto *et al.*, 1990)
23	K1-3	Japan	(Enomoto *et al.*, 1990)
24	K1-4	Japan	(Enomoto *et al.*, 1990)
25	J-121	Japan	This paper
26	HPCGENOM	China	(Bi *et al.*, GenBank number L02836)
27	HPCJTA	Japan	(Tanaka *et al.*, 1992)
28	HPCJTB	Japan	(Tanaka *et al.*, 1992)
29	J483	Japan	(Okamoto *et al.*, 1992*a*)
30	J491	Japan	(Okamoto *et al.*, 1992*a*)
31	HCVJKIG	??	(Honda *et al.*, GenBank number X61596)
32	2TY4	Lebanon	This paper
33	4TY4	Lebanon	This paper
34	K2A	Japan	(Enomoto *et al.*, 1990)
35	SAC640	U.S.A.	This paper
36	HC-J6	Japan	(Okamoto *et al.*, 1991)
37	K2A-1	Japan	(Enomoto *et al.*, 1990)
38	T351	Scotland	This paper
39	T104	Scotland	This paper
40	FC71921	U.S.A.	This paper
41	GC167999	U.S.A.	This paper
42	GC54004	U.S.A.	This paper
43	K2B	Japan	(Enomoto *et al.*, 1990)
44	LQ41461	U.S.A.	This paper
45	K2B-1	Japan	(Enomoto *et al.*, 1990)
46	T59	Scotland	(Chan *et al.*, 1992)
47	T903	Scotland	This paper
48	T810	Scotland	This paper
49	HC-J8	Japan	(Okamoto *et al.*, 1992*b*)
50	ARG6	Argentina	This paper
51	ARG8	Argentina	This paper
52	I10	Italy	This paper
53	T983	Scotland	This paper
54	GH8	Germany	This paper
55	GJ61326	U.S.A.	This paper
56	I11	Italy	This paper
57	I4	Italy	This paper
58	S21	Sweden	This paper
59	T1	Thailand	(Mori *et al.*, 1992)
60	T7	Thailand	(Mori *et al.*, 1992)
61	Eb-1	Scotland	(Chan *et al.*, 1992)
62	Eb-2	Scotland	(Chan *et al.*, 1992)
63	Eb-3	Scotland	(Chan *et al.*, 1992)
64	Eb-7	Scotland	(Chan *et al.*, 1992)
65	T90	Scotland	This paper

Table 1. (*cont.*)

No.	Name	Origin	Reference
66	T1787	Scotland	This paper
67	T9	Thailand	(Mori *et al.*, 1992)
68	T10	Thailand	(Mori *et al.*, 1992)
69	EG-7	Egypt	This paper
70	EG-13	Egypt	This paper
71	EG-19	Egypt	This paper
72	SA156	South Africa	This paper
73	SA183	South Africa	This paper
74	SA30	South Africa	This paper
75	34REV	South Africa	This paper
76	HK-2	Hong Kong	This paper

Nucleotide sequence comparisons. Nucleotide sequences were aligned using the CLUSTAL V program (Higgins *et al.*, 1992) as implemented in the GDE sequence analysis package. Distances between pairs of sequences were estimated using the DNADIST program of the PHYLIP package (version 3.4) kindly provided by Dr J. Felsenstein (Felsenstein, 1991), using a model which allows different rates of transition and transversion and different frequencies of the four nucleotides (Felsenstein, 1991). Phylogenetic trees were constructed using the neighbour-joining algorithm on the previous sets of pairwise distances (Saitou & Nai, 1987) using the PHYLIP program, NEIGHBOR. Equivalent phylogenetic relationships were also found in a maximum likelihood analysis (PHYLIP program DNAML; data not shown), and 2000 bootstrap replicates of neighbour-joining trees (PHYLIP programs SEQBOOT and CONSENSE).

Results and Discussion

Phylogenetic analysis of NS-5 sequences

RNA was extracted from 41 serum or plasma samples from a wide range of HCV-infected individuals from several locations (U.S.A., Europe, South America, South Africa, Middle East and Far East; Table 1), and amplified using previously published NS-5-specific primers (Enomoto *et al.*, 1990). The region was chosen for phylogenetic analysis because it is sufficiently variable to allow differentiation between different isolates of HCV, and there is already a large amount of comparative sequence data; 35 further sequences have been published by other groups and were included in the analysis described here.

Pairwise comparisons revealed a wide range of evolutionary distances amongst the 76 HCV variants analysed. Using Felsenstein's model of molecular evolution (Felsenstein, 1991), distances ranged from 0·01 (between T9 and T10; Mori *et al.*, 1992) to 0·85 (EG-7 to K-2a; Simmonds *et al.*, 1993*a*; Enomoto *et al.*, 1990). However, the distribution of distances was confined to three separate and non-overlapping groups (Fig. 1*a*). The first ranged from 0·38 to 0·84 (mean 0·543), the second showed intermediate evolutionary distances of 0·16 to 0·32 (mean 0·248) and the third showed a range of 0 to 0·12 (mean of 0·061). In none of the 2850 pairwise comparisons were distances of 0·14 to 0·16 or 0·34 to 0·38 found. Furthermore, in no case did the mean value of each distribution ± 3 S.D. overlap with any other, indicating that at least 99·7% of evolutionary distances would be expected to fall within these non-overlapping ranges.

Three levels of sequence diversity may also be observed in a phylogenetic tree of NS-5 sequences (Fig. 2). This analysis shows six major groupings of sequences that are approximately equally divergent from each other. Within some of the major groupings, two or three clusters of more closely related variants are observed.

Mean evolutionary distances between variants in different major branches ranged from 0·41 to 0·66 (Table 2), and distances between the different clusters within the major branches were in all cases substantially lower (0·20 to 0·30). Using the grouping of sequences suggested by Fig. 2, the distribution of mean differences (Table 2) lies entirely within the overall distribution of distances derived from pairwise comparison of individual sequences (Fig. 1).

Based on these results, we propose a classification of HCV that incorporates the three clearly distinct levels of HCV sequence variability. In this scheme, the major groupings of sequence variants are designated HCV 'types', whereas the more closely related groups observed within some types are termed 'subtypes'. In Fig. 2, we have labelled the HCV types in Arabic numbers, and the subtypes by lower case letters, in order of discovery. Thus the genotype of the variant, HCV-1, first cloned by Choo *et al.* (1991) is assigned as type 1a (sequence no. 1).

In this comparison, we have included fragments from a number of other complete genomic sequences. HCV-H (no. 6) shows 95·4% nucleotide sequence identity with HCV-1 over the length of the genome, and is also classified as type 1a in this scheme. Several of the more divergent Japanese and Taiwanese sequences (e.g. HCV-BK, HCV-J and T; nos. 17, 18 and 19) that show 78 to 79% overall sequence similarity with HCV-1 fall into a separate phylogenetic group labelled type 1b in Fig. 2. The most divergent complete genomic sequences HC-J6 and HC-J8 (nos. 36 and 49) that show only 67·1 to 68·3% sequence similarity with types 1a and 1b cluster in separate groups, and have been designated types 2a and 2b respectively.

The group labelled 3a contains NS-5 sequences from variants from Scottish blood donors previously described as type 3 (nos. 61, 62, 63 and 64; Chan *et al.*, 1992), and variants found in NANBH patients from Thailand described by the authors as type V (nos. 59 and 60; Mori *et al.*, 1992). Our phylogenetic analysis of further variants from Thailand originally termed type VI (nos. 67 and 68)

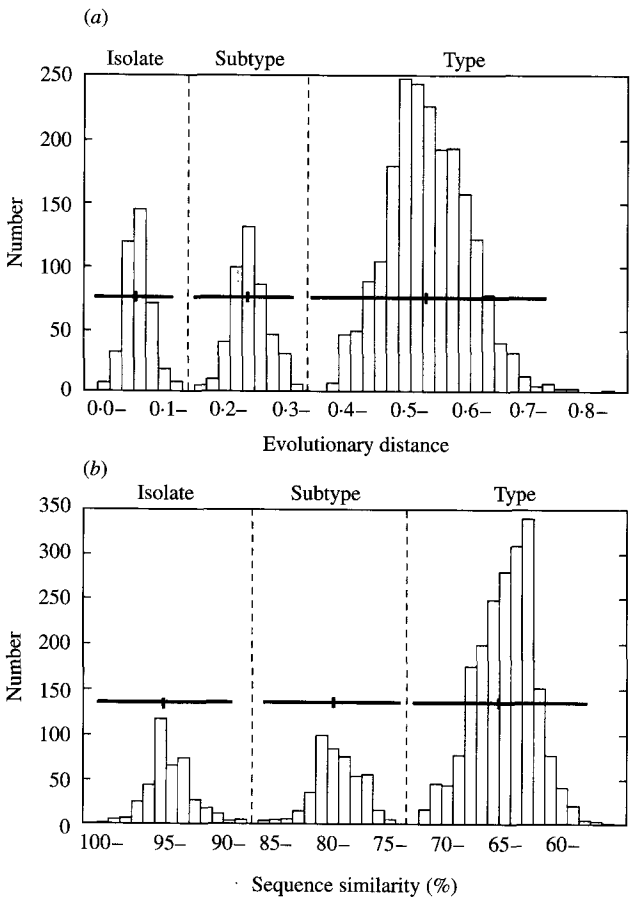

Fig. 1. Distribution of evolutionary distances (*a*) and percentage sequence similarities (*b*) upon pairwise comparison of 76 nucleotide sequences of HCV variants in the NS-5 region (2850 comparisons). (*a*) Number of calculated evolutionary distance measurements (in increments of 0·02) from 0·00 to 0·86 recorded on the *y*-axis. (*b*) Number of observed sequence similarities (in increments of 1 %) recorded on the *y*-axis. Mean ± 3 S.D. for each distribution shown by horizontal bar.

indicated that they were a further subtype of type 3 rather than a major new HCV type.

The six groups, 1a, 1b, 2a, 2b, 3a and 3b, account for all of the previously published NS-5 sequences of HCV and the majority of those obtained in this study. However, from Fig. 2 we can provisionally assign the remainder of the sequences as new types and subtypes of HCV. Sequences amplified from Lebanese NANBH patients (nos. 32 and 33) correspond to a further subtype of type 1, and sequences from NANBH patients in Argentina and Italy, and a Scottish blood donor (nos. 50 to 53) cluster together as a new subtype of type 2.

Variants in the group labelled 4a are all from Egypt. Their status as a new genotype is consistent with our previous analysis of sequences in the core region from these blood donors, where we found that they grouped separately from those of types 1, 2 and 3 (Simmonds *et al.*, 1993*a*). The group labelled 5a contains variants previously designated as group V on the basis of previous sequence analysis of the 5′ NCR and NS-5 regions (Cha *et al.*, 1992). The sequence labelled type 6a (no. 76) was obtained from a Hong Kong blood donor. This variant also differs from all other HCV sequences in the 5′ NCR by the presence of a unique 2 bp insertion at position −143 (Simmonds *et al.*, 1993*a*). Sequences elsewhere in the genome have yet to be examined.

Comparison of the nomenclature shown in Fig. 2 with those originating from other laboratories reveals both similarities and differences (Table 3). Alternative classifications have identified the existence of distinct genotypes, but often do not recognize the two-tiered range of sequence differences. For example, a scheme described by Okamoto *et al.* (1991, 1992*b*) and Mori *et al.* (1992) describes HCV types I, II, III, IV, V and VI which correspond to types 1a, 1b, 2a, 3a and 3b respectively. Cha *et al.* (1992) described five groups of HCV variants, where I corresponds to 1a, II to 1b, III to 2a, 2b and 2c, IV to 3a and 3b to V to type 5a. Consequently, these classifications are difficult to extend to incorporate type 1c, 2c and possibly other new subtypes as they are discovered. Differences between the various schemes have contributed to many of the difficulties in comparing results from different research centres.

Identification of HCV variants

The largest collections of comparative sequence data are currently in the NS-5 region analysed here and in the 5′ NCR. The NS-5 region is a useful region for virus identification because it may be amplified readily from plasma or serum of infected individuals using published primer sequences (Enomoto *et al.*, 1990), and because it is sufficiently variable that both types and subtypes may be identified. Although in principle other parts of the coding region of the viral genome may be equally informative, the current lack of comparative sequence would prevent comparison with the newer sequence variants (types 4 to 6).

Phylogenetic analysis provides the most accurate reconstruction of evolutionary relationships and distances between HCV sequences. However this approach is computationally intensive, and it is easier simply to calculate the proportion of matched nucleotides upon each pairwise comparison of sequences. These (uncorrected) sequence similarities do not allow for multiple substitutions and can greatly underestimate the true extent of divergence between dissimilar sequences (Fig. 1*b*). Sequence similarities between different isolates of the same HCV type (88 to 100%) closely match corresponding evolutionary distances (0 to 0·12), whereas those between subtypes (74 to 86%) and between types (56 to 72%; Fig. 1*b*) differ substantially from the

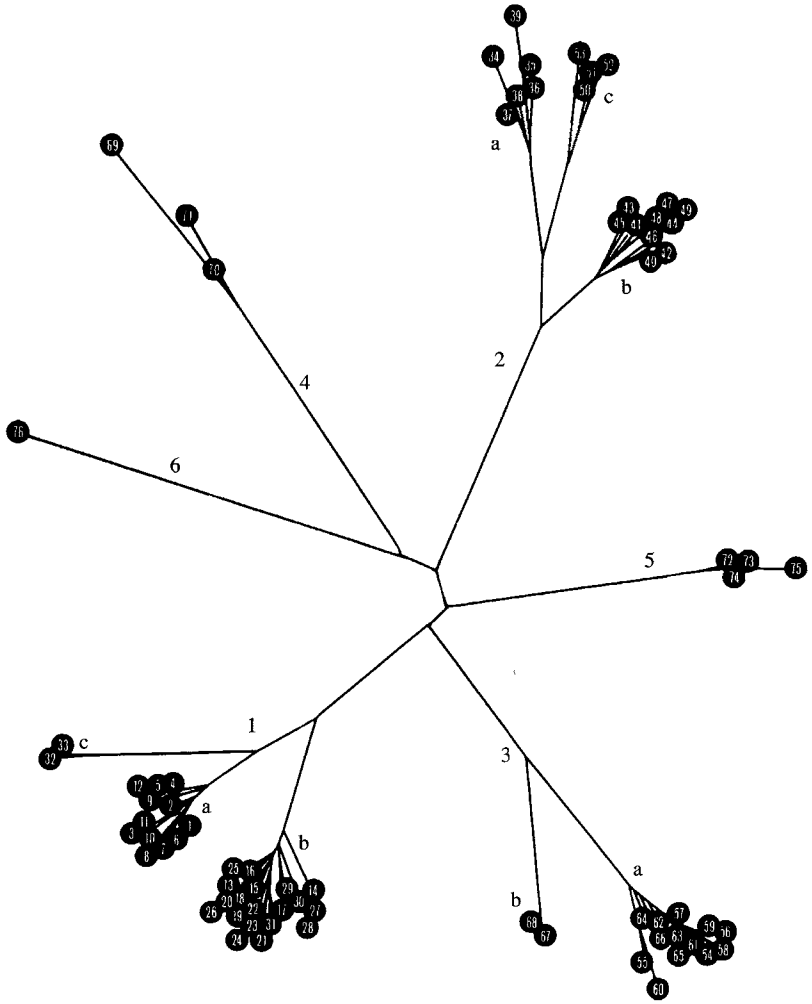

Fig. 2. Phylogenetic analysis of NS-5 sequences from 76 isolates of HCV, showing six major HCV types and subsidiary groupings within some HCV types. Sequences numbered as in Table 1.

Table 2. *Mean evolutionary distances between phylogenetic groupings**

HCV type†	*n*‡	1a	1b	1c	2a	2b	2c	3a	3b	4a	5a	6a
1a	12	**0·0461**										
1b	19	0·2413	**0·0634**									
1c	2	0·1958	0·3017	**0·0465**								
2a	6	0·6263	0·6026	0·5374	**0·0878**							
2b	10	0·5659	0·5177	0·4935	0·2519	**0·0719**						
2c	4	0·6118	0·6022	0·5825	0·2305	0·2577	**0·0892**					
3a	13	0·4926	0·5068	0·5422	0·6392	0·5579	0·6336	**0·0532**				
3b	2	0·5042	0·4207	0·4478	0·6082	0·5079	0·6293	0·2742	**0·0091**			
4a	3	0·5185	0·5967	0·5445	0·6583	0·6154	0·6354	0·5347	0·4578	**0·1354**		
5a	4	0·4535	0·4293	0·4453	0·5502	0·5003	0·6072	0·5132	0·4479	0·5487	**0·0432**	
6a	1	0·5765	0·5289	0·6255	0·5641	0·5388	0·5732	0·5638	0·6234	0·5238	0·4974	NA§

* Distances between subtypes underlined, distances between isolates in bold.
† Groupings derived by phylogenetic analysis of NS-5 sequences (see Fig. 2).
‡ Number of sequences compared within each group.
§ NA, Not applicable.

equivalent ranges for evolutionary distances (0·16 to 0·34 and 0·38 to 0·86; Fig. 1*a*). However, even these uncorrected distances produce three non-overlapping (although compressed) distributions that exactly reproduce the type/subtype distinction derived from evolutionary analysis.

Table 3. *Comparison of nomenclature for HCV types**

Proposed name	Sequence numbers†	Published example	Cha	Chan/ Simmonds	Enomoto	Mori/ Okamoto	Tsukiyama-Kohara
1a	1–12	HCV-1, -H	I	1a	K-PT	I	NC‡
1b	13–31	HCV-J, -BK	II	1b	K-1	II	I
1c	32, 33	—	NC	NC	NC	NC	NC
2a	34–39	HC-J6	III	2a	K-2a	III	II
2b	40–49	HC-J8	III	2b	K-2b	IV	II
2c	50–53	—	III	NC	NC	NC	NC
3a	54–66	Ta, E-b1	IV	3	NC	V	NC
3b	67, 68	Tb	IV	NC	NC	VI	NC
4a	69–71	—	NC	4	NC	NC	NC
5a	72–75	—	V	NC	NC	NC	NC
6a	76	—	NC	NC	NC	NC	NC

* Proposed nomenclature for published HCV sequences and comparison with existing schemes Cha: Cha *et al.* (1992); Chan/Simmonds: Chan *et al.* (1992); Simmonds *et al.* (1993*a*); Enomoto: Enomoto *et al.* (1990); Mori/Okamoto: Okamoto *et al.* (1992*b*); Mori *et al.* (1992); Tsukiyama-Kohara: Tsukiyama Kohara *et al.* (1991).

† Sequences numbered as in Table 1.

‡ NC, Sequences not classified by originating authors.

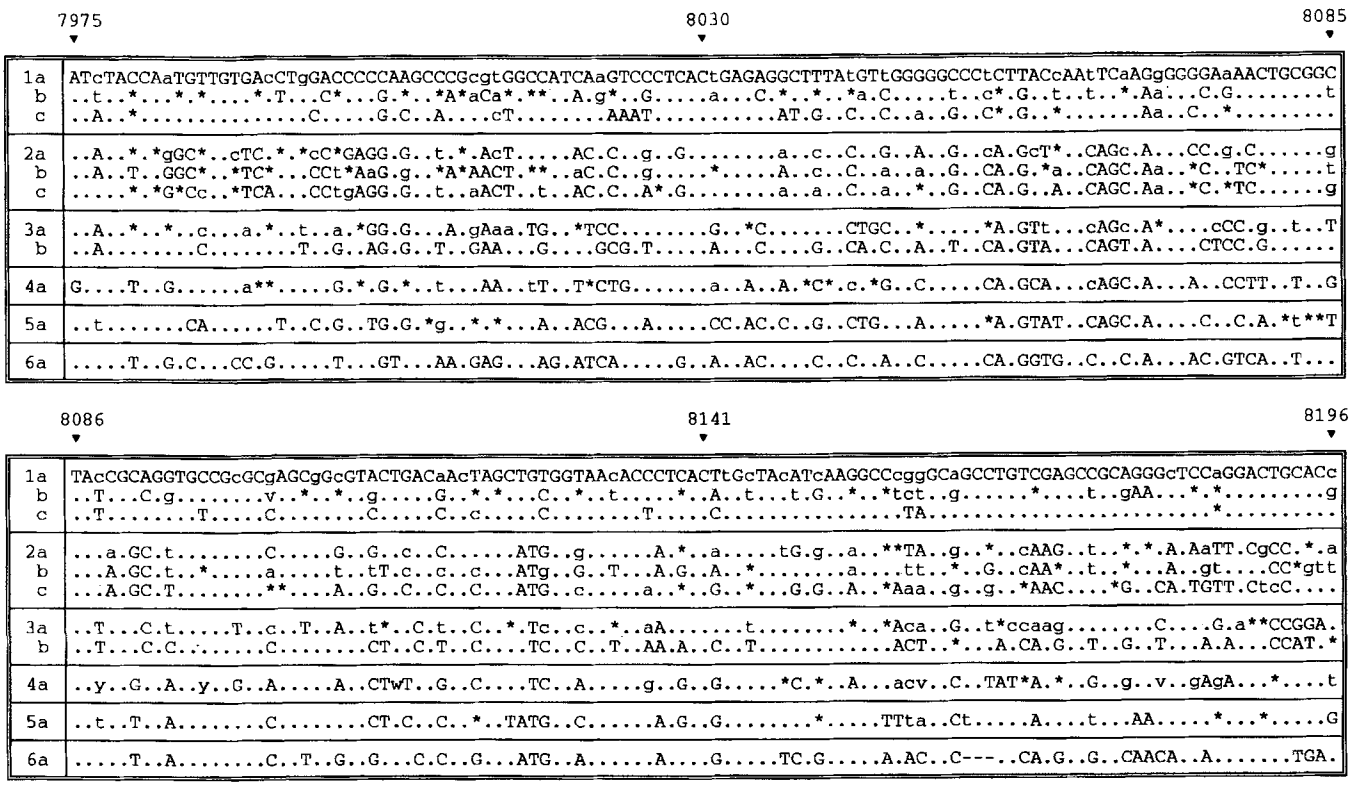

Fig. 3. Comparison of consensus nucleotide sequences of part of NS-5 from HCV types 1 to 6. Numbers of sequences assigned to each type and subtype are shown in Table 2. Lower case letters indicate variable sites within sequence group, majority nucleotide shown. Where no nucleotide forms a majority, IUPAC ambiguity codes are used: 'y', C or T; 'v', G, C or A; 'w', T or A. Other symbols: dots (·), identity with type 1a sequence; hyphens represent gap introduced to preserve sequence alignment (type 6a); asterisks represent variable nucleotide positions where majority sequence is the same as type 1a. Sequences numbered as in Choo *et al.* (1991).

The consensus sequences of each type and subtype described in this paper (Fig. 3) represent an 'average' sequence for each HCV group and may be conveniently used for the identification and provisional classification of new HCV variants. Sequence similarities of less than 72% with any of the 11 consensus nucleotide sequences indicate that the new sequence should be designated a new HCV type. Those showing sequence similarities of

between 75% and 86% with particular variants and less than 72% with other should be assigned as a new subtype. Clearly, if the new sequence showed greater than 88% similarity with any of the consensus sequences in Fig. 3, then it should not be assigned a new type description. It should be stressed that this type of classification must be regarded as provisional, and a more complete analysis of evolutionary relationships by phylogeny would be required to confirm sequence designations.

Sequence comparisons in other regions of the genome

We have carried out comparable analysis in the core region using sequences of HCV types 1 to 4 (Simmonds *et al.*, 1993*a*). This region also produces three non-overlapping distributions of sequence similarities, although in this case, different HCV types show much greater similarities than for NS-5 (data not shown). Using the region of core shown in our previous analysis and sequences listed therein (positions 29 to 269; Simmonds *et al.*, 1993*a*), we found similarities between HCV types ranged from 81 to 89% whereas those between subtypes were from 91 to 94%. Sequence similarities of greater than 94% are found between different isolates of the same HCV type.

All of the six major types of HCV may also be identified by comparison of sequences in the 5′ NCR (Simmonds *et al.*, 1993*a*; Bukh *et al.*, 1992*b*; Cha *et al.*, 1992), although often only relatively few nucleotide differences exist between them. Nucleotide sequence variability in the 5′ NCR also differs from that found in the coding regions in that there are no reliable sequence polymorphisms in this region that allow the differentiation of HCV subtypes. HCV types 1a, 1b and 1c show essentially the same sequences in the 5′ NCR, as do types 2a and 2c (Chan *et al.*, 1992; Cha *et al.*, 1992; Simmonds *et al.*, 1993*a*). At this stage, it appears that reliable differentiation into subtypes will remain dependent on analysis of coding regions.

Origins of HCV types

The samples used in this study were from either blood donors or patients with non-A, non-B hepatitis. Although a comprehensive geographical survey of HCV types was not attempted, there do appear to be differences in their distribution in different countries. HCV types 1 and 2 have been found in almost all countries tested, including those in Europe, North America and the Far East (Li *et al.*, 1991; Takada *et al.*, 1992*a*, *b*; Cha *et al.*, 1992; Kato *et al.*, 1991; Chan *et al.*, 1992). Although HCV type 3 has not been found in Japan, it has been frequently reported from Europe (Chan *et al.*, 1992), U.S.A. (Lee *et al.*, 1992), Thailand (Mori *et al.*, 1992) and India (unpublished observations). In the Middle East, almost all anti-HCV-positive individuals identified on blood donor screening are infected with type 4 (Simmonds *et al.*, 1993*a*), which has also been detected in NANBH patients in Zaire (Bukh *et al.*, 1992*b*). HCV types 5 and 6 show highly restricted geographical distributions, being apparently confined to South Africa and Hong Kong respectively (Cha *et al.*, 1992; Simmonds *et al.*, 1993*a*). At this stage it is difficult to interpret the significance of these different distributions of variants, as the epidemiology of virus spread is at present poorly understood.

In summary, we have used a series of comparative methods to establish sequence relationships between HCV variants. The system is internally consistent, and the phylogenetic and numerical comparison methods described here will facilitate the assigment of new sequence variants as they are discovered. The uniform nomenclature proposed in Table 2 would, if adopted, considerably clarify comparative evaluation of results from different laboratories. This requirement will undoubtedly increase as more is understood about the important biological and serological differences that have been found to exist between the different variants of HCV.

The authors would like to acknowledge Dr E. A. C. Follett and staff at the Scottish National Blood Transfusion Service, Dr A. A. Saaed, Riyadh Armed Forces Hospital, Riyadh, Saudi Arabia, and Drs C. K. Lin, S. Leong and C. Lai, Hong Kong Red Cross Blood Transfusion Service, Kowloon, Hong Kong for providing samples from HCV-infected blood donors for sequence analysis. We are grateful to Donald Smith for careful reading of the paper for submission and the mainly helpful comments he made. The work was funded in part by the Medical Research Council (S.-W. Chan; grant number G9020615CA) and the Scottish National Blood Transfusion Service (F. McOmish). Work carried out at Chiron was funded in part by Daiichi Pure Chemicals Co. Ltd., Tokyo, Japan.

References

Brown, E. A., Zhang, H., Ping, H.-L. & Lemon, S. M. (1992). Secondary structure of the 5′ nontranslated region of hepatitis C virus and pestivirus genomic RNAs. *Nucleic Acids Research* **20**, 5041–5045.

Bukh, J., Purcell, R. H. & Miller, R. H. (1992*a*). Importance of primer selection for the detection of hepatitis C virus RNA with the polymerase chain reaction assay. *Proceedings of the National Academy of Sciences, U.S.A.* **89**, 187–191.

Bukh, J., Purcell, R. H. & Miller, R. H. (1992*b*). Sequence analysis of the 5′ noncoding region of hepatitis C virus. *Proceedings of the National Academy of Sciences, U.S.A.* **89**, 4942–4946.

Cha, T. A., Kolberg, J., Irvine, B., Stempien, M., Beall, E., Yano, M., Choo, Q. L., Houghton, M., Kuo, G., Han, J. H. & Urdea, M. S. (1991). Use of a signature nucleotide sequence of hepatitis C virus for detection of viral RNA in human serum and plasma. *Journal of Clinical Microbiology* **29**, 2528–2534.

Cha, T. A., Beall, E., Irvine, B., Kolberg, J., Chien, D., Kuo, G. & Urdea, M. S. (1992). At least five related, but distinct C viral genotypes exist. *Proceedings of the National Academy of Sciences, U.S.A.* **89**, 7144–7148.

CHAN, S.-W., SIMMONDS, P., MCOMISH, F., YAP, P.-L., MITCHELL, R., DOW, B. & FOLLETT, E. (1991). Serological reactivity of blood donors infected with three different types of hepatitis C virus. *Lancet* **338**, 1391.

CHAN, S.-W., MCOMISH, F., HOLMES, E. C., DOW, B., PEUTHERER, J. F., FOLLETT, E., YAP, P. L. & SIMMONDS, P. (1992). Analysis of a new hepatitis C virus type and its phylogenetic relationship to existing variants. *Journal of General Virology* **73**, 1131–1141.

CHEN, P. J., LIN, M. H., TAI, K. F., LIU, P. C., LIN, C. J. & CHEN, D. S. (1992). The Taiwanese hepatitis C virus genome: sequence determination and mapping the 5′ termini of viral genomic and antigenomic RNA. *Virology* **188**, 102–113.

CHOO, Q. L., KUO, G., WEINER, A. J., OVERBY, L. R., BRADLEY, D. W. & HOUGHTON, M. (1989). Isolation of a cDNA derived from a blood-borne non-A, non-B hepatitis genome. *Science* **244**, 359–362.

CHOO, Q. L., RICHMAN, K. H., HAN, J. H., BERGER, K., LEE, C., DONG, C., GALLEGOS, C., COIT, D., MEDINA SELBY, R., BARR, P. J., WEINER, A. J., BRADLEY, D. W., KUO, G. & HOUGHTON, M. (1991). Genetic organization and diversity of the hepatitis C virus. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 2451–2455.

ENOMOTO, N., TAKADA, A., NAKAO, T. & DATE, T. (1990). There are two major types of hepatitis C virus in Japan. *Biochemical and Biophysical Research Communications* **170**, 1021–1025.

FELSENSTEIN, J. (1991). *PHYLIP Manual Version* 3.4. Berkeley: University Herbarium, University of California.

HAN, J. H., SHYAMALA, V., RICHMAN, K. H., BRAUER, M. J., IRVINE, B., URDEA, M. S., TEKAMP OLSON, P., KUO, G., CHOO, Q. L. & HOUGHTON, M. (1991). Characterization of the terminal regions of hepatitis C viral RNA: identification of conserved sequences in the 5′ untranslated region and poly(A) tails at the 3′ end. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 1711–1715.

HIGGINS, D. G., BLEASBY, A. J. & FUCHS, R. (1992). ClustalV: improved software for multiple sequence alignments. *CABIOS* **8**, 189–191.

HIJIKATA, M., KATO, N., OOTSUYAMA, Y., NAKAGAWA, M., OHKOSHI, S. & SHIMOTOHNO, K. (1991). Hypervariable regions in the putative glycoprotein of hepatitis C virus. *Biochemical and Biophysical Research Communications* **175**, 220–228.

INCHAUSPE, G., ZEBEDEE, S., LEE, D. H., SUGITANI, M., NASOFF, M. & PRINCE, A. M. (1991). Genomic structure of the human prototype strain H of hepatitis C virus: comparison with American and Japanese isolates. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 10292–10296.

KANAI, K., KAKO, M. & OKAMOTO, H. (1992). HCV genotypes in chronic hepatitis-C and response to interferon. *Lancet* **339**, 1543.

KATO, N., HIJIKATA, M., OOTSUYAMA, Y., NAKAGAWA, M., OHKOSHI, S., SUGIMURA, T. & SHIMOTOHNO, K. (1990). Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. *Proceedings of the National Academy of Sciences, U.S.A.* **87**, 9524–9528.

KATO, N., OOTSUYAMA, Y., OHKOSHI, S., NAKAZAWA, T., MORI, S., HIJIKATA, M. & SHIMOTOHNO, K. (1991). Distribution of plural HCV types in Japan. *Biochemical and Biophysical Research Communications* **181**, 279–285.

KUBO, Y., TAKEUCHI, K., BOONMAR, S., KATAYAMA, T., CHOO, Q. L., KUO, G., WEINER, A. J., BRADLEY, D. W., HOUGHTON, M., SAITO, I. & MIYAMURA, T. (1989). A cDNA fragment of hepatitis C virus isolated from an implicated donor of post-transfusion non-A, non-B hepatitis in Japan. *Nucleic Acids Research* **17**, 10367–10372.

KUO, G., CHOO, Q. L., ALTER, H. J., GITNICK, G. L., REDEKER, A. G., PURCELL, R. H., MIYAMURA, T., DIENSTAG, J. L., ALTER, M. J., STEVENS, C. E., TEGTMEIER, F., BONINO, F., COLUMBO, M., LEE, W.-S., KUO, C., BERGER, K., SCHUSTER, J. R., OVERBY, L. R., BRADLEY, D. W. & HOUGHTON, M. (1989). An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. *Science* **244**, 362–364.

LEE, C., CHENG, C., WANG, J. & LUMENG, L. (1992). Identification of hepatitis C viruses with a nonconserved sequence of the 5′ untranslated region. *Journal of Clinical Microbiology* **30**, 1602–1604.

LI, J., TONG, S., VITVITSKI, L., LEPOT, D. & TREPO, C. (1991). Two French genotypes of hepatitis C virus: homology of the predominant genotype with the prototype American strain. *Gene* **105**, 167–172.

MCOMISH, F., CHAN, S.-W., DOW, B. C., GILLON, J., FRAME, W. D., CRAWFORD, R. J., YAP, P. L., FOLLETT, E. A. C. & SIMMONDS, P. (1993). Detection of three types of hepatitis C virus in blood donors: investigation of type-specific differences in serological reactivity and rate of alanine aminotransferase abnormalities. *Transfusion* **33**, 7–13.

MORI, S., KATO, N., YAGYU, A., TANAKA, T., IKEDA, Y., PETCHCLAI, B., CHIEWSILP, P., KURIMURA, T. & SHIMOTOHNO, K. (1992). A new type of hepatitis C virus in patients in Thailand. *Biochemical and Biophysical Research Communications* **183**, 334–342.

OGATA, N., ALTER, H. J., MILLER, R. H. & PURCELL, R. H. (1991). Nucleotide sequence and mutation rate of the H strain of hepatitis C virus. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 3392–3396.

OKAMOTO, H., OKADA, S., SUGIYAMA, Y., YOTSUMOTO, S., TANAKA, T., YOSHIZAWA, H., TSUDA, F., MIYAKAWA, Y. & MAYUMI, M. (1990). The 5′-terminal sequence of the hepatitis C virus genome. *Japanese Journal of Experimental Medicine* **60**, 167–177.

OKAMOTO, H., OKADA, S., SUGIYAMA, Y., KURAI, K., IIZUKA, H., MACHIDA, A., MIYAKAWA, Y. & MAYUMI, M. (1991). Nucleotide sequence of the genomic RNA of hepatitis C virus isolated from a human carrier: comparison with reported isolates for conserved and divergent regions. *Journal of General Virology* **72**, 2697–2704.

OKAMOTO, H., KOJIMA, M., OKADA, S.-I., YOSHIZAWA, H., IIZUKA, H., TANAKA, T., MUCHMORE, E. E., ITO, Y. & MISHIRO S. (1992*a*). Genetic drift of hepatitis C virus during an 8·2 year infection in a chimpanzee: variability and stability. *Virology* **190**, 894–899.

OKAMOTO, H., KURAI, K., OKADA, S., YAMAMOTO, K., IIZUKA, H., TANAKA, T., FUKUDA, S., TSUDA, F. & MISHIRO, S. (1992*b*). Full-length sequence of a hepatitis C virus genome having poor homology to reported isolates: comparative study of four distinct genotypes. *Virology* **188**, 331–341.

POZZATO, G., MORETTI, M., FRANZIN, F., CROCE, L. S., TIRIBELLI, C., MASAYU, T., KANEKO, S., UNOURA, M. & KOBAYASHI, K. (1991). Severity of liver disease with different hepatitis C viral clones. *Lancet* **338**, 509.

SAITOU, N. & NEI, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406–425.

SIMMONDS, P. & CHAN, S.-W. (1993). Analysis of viral sequence variation by PCR. In *Molecular Virology: A Practical Approach*, pp. 109–138. Edited by A. J. Davison & R. M. Elliott. Oxford: IRL Press.

SIMMONDS, P., BALFE, P., LUDLAM, C. A., BISHOP, J. O. & BROWN, A. J. (1990). Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *Journal of Virology* **64**, 5840–5850.

SIMMONDS, P., MCOMISH, F., YAP, P. L., CHAN, S.-W., LIN, C. K., DUSHEIKO, G., SAEED, A. A. & HOLMES, E. C. (1993*a*). Sequence variability in the 5′ non-coding region of hepatitis C virus: identification of a new virus type and restrictions on sequence diversity. *Journal of General Virology* **74**, 661–668.

SIMMONDS, P., ROSE, K. A., GRAHAM, S., CHAN, S. W., MCOMISH, F., DOW, B. C., FOLLETT, E. A. C., YAP, P. L. & MARSDEN, H. (1993*b*). Mapping of serotype-specific, immunodominant epitopes in the NS-4 region of hepatitis C virus (HCV) – use of type-specific peptides to serologically differentiate infections with HCV type 1, type 2, and type 3. *Journal of Clinical Microbiology* **31**, 1493–1503.

TAKADA, N., TAKASE, S., ENOMOTO, N., TAKADA, A. & DATE, T. (1992*a*). Clinical backgrounds of the patients having different types of hepatitis C virus genomes. *Journal of Hepatology* **14**, 35–40.

TAKADA, N., TAKASE, S., TAKADA, A. & DATE, T. (1992*b*). HCV genotypes in different countries. *Lancet* **339**, 808.

TAKAMIZAWA, A., MORI, C., FUKE, I., MANABE, S., MURAKAMI, S., FUJITA, J., ONISHI, E., ANDOH, T., YOSHIDA, I. & OKAYAMA, H. (1991). Structure and organization of the hepatitis C virus genome isolated from human carriers. *Journal of Virology* **65**, 1105–1113.

TANAKA, T., KATO, N., NAKAGAWA, M., OOTSUYAMA, Y., CHO, M. J., NAKAZAWA, T., HIJIKATA, M., ISHIMURA, Y. & SHIMOTOHNO, K. (1992). Molecular cloning of hepatitis C virus genome from a single Japanese carrier: sequence variation within the same individual and among infected individuals. *Virus Research* **23**, 39–53.

TSUKIYAMA KOHARA, K., KOHARA, M., YAMAGUCHI, K., MAKI, N.,

TOYOSHIMA, A., MIKI, K., TANAKA, S., HATTORI, N. & NOMOTO, A. (1991). A second group of hepatitis C viruses. *Virus Genes* **5**, 243–254.

WEINER, A. J., BRAUER, M. J., ROSENBLATT, J., RICHMAN, K. H., TUNG, J., CRAWFORD, K., BONINO, F., SARACCO, G., CHOO, Q. L., HOUGHTON, M. & HAN, J. H. (1991). Variable and hypervariable domains are found in the regions of HCV corresponding to the flavivirus envelope and NS1 proteins and the pestivirus envelope glycoproteins. *Virology* **180**, 842–848.

YANISCH-PERRON, C., VIEIRA, J. & MESSING, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**, 103–107.

YOSHIOKA, K., KAKUMU, S., WAKITA, T., ISHIKAWA, T., ITOH, Y., TAKAYANAGI, M., HIGASHI, Y., SHIBATA, M. & MORISHIMA, T. (1992). Detection of hepatitis C virus by polymerase chain reaction and response to interferon-alpha therapy: relationship to genotypes of hepatitis C virus. *Hepatology* **16**, 293–299.

(*Received 17 August 1993; Accepted 19 August 1993*)